12-5-2021

Using Prediction ML algorithm for predicting early Student Attrition in Higher Education

Zainab AlHashemi
za6346@rit.edu

RIT

Using Prediction ML algorithm for predicting early Student Attrition in Higher Education

by

Zainab AlHashemi

A Capstone Submitted in Partial Fulfilment of the Requirements for the Degree of Master of Science in Professional Studies:

Data Analytics

Department of Graduate Programs & Research

Rochester Institute of Technology

RIT Dubai

5 December 2021

RIT

Master of Science in Professional Studies:

Data Analytics

Graduate Capstone Approval

Student Name: Zainab AlHashemi

Graduate Capstone Title: Using Prediction ML algorithm for predicting early Student Attrition in Higher Education

Graduate Capstone Committee:

Name: Dr. Sanjay Modak **Date:**

Chair of committee

Name: Dr. Ehsan Warriach **Date:**

Member of committee

Acknowledgments

I’d like to express my deepest thanks to my mentor Dr. Ehsan for his support and guidance during my work on the project. I would like to also thank my group members for their positive words and encouragement during my studies. Although I have little knowledge in languages such as R and Python, I’m appreciating every knowledge I received at Rochester Institute of Technology Dubai and the answers to the questions I received from every instructor. Lastlly, I would like to show my appreciation to my family and friends for believing in me and pushing me to achieve my goals. Lastly, to my children, thank you for understanding.

Abstract

This research aims at using predictive models that enable us to predict students who are at risk of dropping out and identify the factors that possibly lead to this dropout. Through the results obtained, concerned stakeholders will be able to effectively develop strategies and initiatives to help decrease the percentage of students' attrition. There are different reasons why students drop from their courses which could be related to academic issues or personal issues that stop them from being active students. Due to these many reasons of students dropping out, universities are impacted negatively in terms of the financial costs as they lose an amount of money from those students, and sometimes they lose the funds from public sponsors to major activities in universities. The proposal aims at exploring the various reasons that influence students' decision to withdraw and what will be the best model for the prediction. I will use data from the open-source Kaggle and use Python to explore and preprocess the data. I will also use Tableau for getting visual insights from the available dataset.

Keywords: Attrition, Higher Education, Dropout, Students, Higher Education Institutions, Completion Rate, Return, Advising, Dropout Risk

Table of Contents

List of Figures

List of Tables

Chapter 1

1.1 Background

Generally, the increase in student attrition rates in higher education institutions has negative effects on both the individual and the institution. In recent years in particular due to the pandemic and online learning, it has become very important for institutions to understand the factors that influence students to withdraw from higher education. Due to the implications students' withdrawal has on the institutions in terms of reputation and revenue, it has become a risk factor for them to search for the reasons and come up with recommendations to solve this issue.

In this report, previous work in regards to the students' attrition rates and the different models used for early prediction were reviewed. As a finding, it turns that different factors affect students' decision to continue studying at the university or not through which the institutions can develop strategies to address this issue before it occurs. Many think negatively about withdrawn students in terms they are lower performers, fail a lot, or even have high absence rates, however, this might not be true as there are non-academic factors that influence the decision to drop out of a university. Additionally, many of these students gain skills that they didn't have before they join a university.

A predictive model will be built based on the dataset which will be available for key stakeholders to use for developing correction action plans to decrease the number of attrition.

1.2 Problem Statement

Student attrition is defined as decreasing number of students who study in the university. It counts the number of students who drop out of their courses before they graduate. Student attrition is one of the main issues that have a significant impact on higher education institutions in terms of costs and reputation. Some agencies measure the performance of universities through student attrition rates; hence, this is one of the important factors that universities try putting efforts into decreasing the number of students who drop out. Hence, it is not only important to know the reasons for students drop-out, but more importantly, it is important to be able to identify those students who might turn to drop out.

1.3 Project Goals

There are two main goals for this project: (a) Identify the factors that influence students' decision to withdraw from a university (b) Explore models that would accurately predict those students who are at risk of dropping out. Additionally, I shall explore different journal articles to understand the different techniques used by the authors for such an issue and what were the limitations in their studies.

Research questions

What are the reasons behind students' choices to drop from a university?

What triggers students to withdraw from the university?

What is the best model to use in our prediction?

1.4 Methodology

To be able to achieve the best results, the best approach to use is the CRISP-DM method by following the below steps:

Stage 1: Business Understanding

Where we identify the potential causes for students' choices to withdraw from a university and use the findings in reducing the attrition rates in the educational institutions that are affected by this issue financially as well as through their reputation. There could be many factors that influence this issue where each factor might affect a specific group of students. Once the factors are identified, it will be easier for the business to concentrate more on those students who are highly affected by these factors.

Stage 2: Data Understanding

In this phase, all angles related to the data will be explored such as the source of the data, the description of the data, and the quality dimensions of the dataset collected. What is more, data exploration will include identifying the types of the variables, the null values available, and other features. This phase is also called the Exploratory Data Analysis.

Stage 3: Data Preparation

In order to work with the dataset from the previous section, many actions need to be taken to clean that data in particular so we can build a proper model.

1. Identifying the fields that include the null values and filling those as appropriate
2. Dropping unnecessary variables in case they do not add any value to the project
3. Dropping months/years from dates in the variables if they add no value

4. Retrieving new variables from existing variables wherever necessary

Stage 4: Modeling

In this phase, I will be using Python to develop models to predict students' attrition based on the dataset using different machine-learning algorithms. Additionally, for the visualizations, Tableau will be used along with Python to generate visual insights.

Stage 5: Evaluation

The final phase is about comparing the different models or the model with different comparison matrixes to choose the best model in case there is more than one. For this project, I'm intending to predict students' attrition using different comparison matrices such as accuracy, sensitivity, and AUC on the test dataset.

1.5 Limitations of the Study

There were few limitations when working with this project:

- The entity did not share a dataset to work with due to confidentiality and data security-related issues. Hence, had to search for a proper dataset to start working on the project
- In the dataset, there are a lot of attributes with null values where I had doubts when starting to work on the dataset
- The dataset consisted of 56 attributes where many of which did not seem to be beneficial
- The last update the dataset has received was before two years. Having recent data might have given accurate data to work with

Chapter 2 – Literature Review

In this part, I shall go through different journal articles that highlight the issue of students' withdrawal and the factors that influence their decisions for dropping out. What is more, I will also look at the different prediction techniques used by the authors in their articles to find the limitations in their studies and try to find a solution that will add value to the previous researches.

According to Lee and Chung (2019), students drop out not only affects the students directly but can also affect society in different ways. For the students, their drop-out will impact their well-being in the later lives where they will miss learning new knowledge and the skills which they can use later in their professional lives. As for society, the more students drop out, the society will not have the qualified people to work for its organizations, and also will increase the unemployment rates. In their research, both Lee and Chung tried to forecast early student dropout by using different techniques such as synthetic as well as trained classifiers of random forest (RF), boosted decision tree (BDT), random forest with SMOTE (SMOTE + RF), and boosted decision tree with SMOTE (SMOTE + BDT). The findings were not very much informative for reasons such as not all information being accessible to do further analysis. Hence, the authors stated that further future studies are needed to analyze student attrition through class imbalance.

Aulck et al (2016) stated that around 30% of students in the first year don’t return for their second year in US higher education where over $9 billion is spent to educate these students. The authors used the largest dataset on higher education which tracks students’ demographics and academic records in one of the largest universities. They’ve used this dataset to develop a model that predicts students’ dropouts even if the dataset belonged to one term. The results showed GPA in math, English, Chemistry, and Psychology courses as main predictors for student dropouts. Additionally,

marginal results showed when the authors wanted to analyze the number of quarters taken before dropping out. To enhance the results better, there are intentions to speak to other university administrations and get datasets for universities where the student attrition rate is high.

According to Grau-Valldosera & Minguillón (2014), they have highlighted that the factors that affect students' decisions to drop from traditional on-campus programs than the online are different. Hence, they've identified student attrition as those students who join a university during their life span, however, they fail to complete it.

According to Shaw et al (2016), students' withdrawal is higher in the online institutions 5% more than the traditional ones. Due to this, they have designed a tracking method to track the factors that influence the attrition in the online courses and of those identifying the students affected by these factors. This helped them to develop retention initiatives for the at-risk students to withdraw. They have used several methods to determine the critical factors that contributed to the students not completing their studies such as quantitative, experimental, and correlational designs methods. The findings showed that students who study verbal learning styles are more likely to drop than those who study programs that add to their skills. What is more, students who put their studies on hold due to personal reasons so they can continue the studies the following semester are also prone to drop eventually. As for recommendations, the authors recommended having this tracking system that they've designed to help track the students at risk of withdrawal so more support activities could be proposed for them to help them overcome the difficulties.

According to AlJohani (2016), his studies showed that those students who lack the knowledge of different types of institutions' offerings in terms of academic advancement and career opportunities are more likely to drop than those with sufficient knowledge. His studies covered attrition in four-year and two-year higher education institutions. For those students who are not familiar with

institutions, they might end up with wrong decisions in terms of transferring to other institutions that they might think will help them academically or increase their potentials professionally, or even choosing to withdraw to be able to work.

As Maher & Macallister (2013) stated in their journal article, there are a lot of existing articles that wrote about the different effects of student attrition in higher education institutions. Those effects not only affect the institutions, but also the students and the staff on a personal level. The main goal of their paper is to find out the main factors for students' retention which can also be applied to higher educational institutions. The authors used mixed methods where the data collection consisted of statistics about the rates of students' attrition and retention, also, they've conducted interviews with the relevant employees. The conclusion of their article is that both individuals and institutions face financial losses as well as limited return on investment when students drop.

In a journal article by Johnson (2012), he states that there are financial implications on the higher education institution due to students' drop out that his findings highlight that about 35% of students discontinue their students before they complete the full academic plan. He indicates that it doesn't matter whether the students chose to transfer to other less expensive institutions or entirely gave up on their studies, the implications on the universities are the same. According to Johnson, universities spend on an individual student to the amount of $43K and around $18K on those to withdraw which indicates that because of these costs on the universities, the facilities are not utilized at a full capacity which puts the institution at a financial loss with no profit. As cited in Johson (2012), Schneider stated that taxpayers, as well as the state, spend an amount close to $9 billion on students who might choose to withdraw the following year. Due to this, the revenues the educational institutions get from the state and the taxpayers are put in jeopardy as they depend mostly on these funds to run their educational institutions.

In a journal article by Abu Oda & ElHalees (2015), they were aiming to identify those students who are less likely to return from one semester to another from a computer science program. The authors used data collected from Al-Aqsa university for bachelor students with records of 1290 representing the students and their transcripts. They used different classification techniques to predict and examine students' dropouts throughout the study. Of those methods, Decision tree (DT) and Naïve Bayes (NB) techniques. The results showed that mastering courses such as digital design and algorithm analysis have a great effect on predicting students' continuity in the program and decrease the chances of dropouts.

To successfully predict student attrition, Berens et al (2019), developed an Early Detection System (EDS) using data from private and state universities. Instead of relying only on one method, different techniques were used to build the models such as AdaBoost Algorithm to combine regression analysis, neural networks, and decision trees. The prediction of the accuracy was done in two phases, at the end of the first semester and after the fourth semester. The results showed that the accuracy of the data increases through time, meaning that the accuracy increases in the fourth semester every time the model calculates the accuracy. One of the limitations is that the available demographic data is only relevant to early detection in the first year because, by the time, the model also reads from the available performance data.

In an article by Yukselturk et al (2014), the main aim was to predict student dropout in an online program using different data mining algorithms. The authors used 189 records of an online certification program in information technology. They've collected the data using an online questionnaire where the data collected included ten attributes such as demographics, current, and past educational experience, and the dropout status as the class label. To successfully classify student dropout, the authors used four classification approaches: K nearest neighbor, decision tree,

naïve bayes, and neural network. Although there were no significant differences in the sensitivities among the four, the K nearest neighbors and decision tree were more sensitive which showed alignment in results with previous research in online programs. The main limitations were the single experiment and the limited number of sample data where future researchers could collect more data, use other variables, and apply other algorithms.

In a case study conducted by Dekker et al (2009) to predict student dropout in the first year of the university, the authors considered data from 2000 – 2009 that consists of students registered in the target program, Electrical Engineering. The authors considered three datasets: pre-university data, university grades only, and a dataset that's containing both variables. The class selected for the prediction was "successful" and "unsuccessful". To predict, the authors used Weka's built-in classifier models such as decision tree, Bayesian classifier, logistic model, a rule-based learner, and random forest. There were several techniques used to check the accuracy of the models such as cost sensitivity and accuracy. A few limitations were that the test model needed to be improved due to the fact that there were real differences between different model classifiers in their results. There were 25% of misclassified instances.

To be able to evaluate the performance and dropouts of undergraduates, Manhães et al (2014) used Educational Data Mining (EDM) for one of the Brazilian universities. According to Baker and Yacef (2009), EDM is a recent research area concerned with using computerized methods to detect patterns in enormous educational data that is difficult/impossible to analyze manually. The solution that Manhães et al developed was a multi-tier architecture model where it includes analytical functionalities. Classifications algorithms such as Naïve Bayes (NB), Multilayer Perceptron (MLP), Support Vector Machine (SVM) with polynomial kernel and RBF kernel, and Decision Tree (DT) were used through Weka. To investigate, six undergraduate courses were

used in the evaluation because they fit the criteria of the assessment and they include different dropout rates. The test showed an accuracy of over 70% that students will drop out. The authors described their architecture as one of the first to use all internal data where no external data is required.

Student dropout issues are a concern for many universities for the financial and reputations impact they have. Hence, many researchers try to find out the reasons behind students' attrition using analytical tools to be able to deal with the enormous amount of data. Different machine learning algorithms have been used in classifications to predict students with high accuracy of dropping from universities. As a learning lesson, in order to find the best classification model to predict students' attrition, we have to try different models and use the comparison metrics between these models. Once we have the best model, we need to also review the model once in a while as data changes. What is more, there are new tools used for classifications, other than the machine learning programs such as Python and R, there's also Weka which is used by educators. The limitations of all the studies, included the availability of data, tools to deal with the enormous amount of data where one of the authors developed an architecture tool, and finding the best model with high accuracy.

Chapter 3- Project Description

To be able to work with the collected dataset, this project will go through many steps to find the best model. First, the dataset will go through a preprocessing stage where the dataset is transformed from raw data to a format that allows us to understand the dataset. This is considered an important step in data mining to find knowledge and work through different programming languages and tools to get insights. Through this stage, data cleaning techniques will be utilized such as filling the null values. Also, select the proper attributes that help us in our project rather than working with all attributes. Once the data is clean and ready to get some information, the next stage will be is to get insights through visualizations. Lastly, use some algorithms to help us in modeling to get the outcomes we aimed at in this project.

3.1 Data Collection

Many open data sources were looked at in order to find the relevant dataset to work with including Dubai Pulse, US Census Bureau, Data.Gov which is considered the treasure-house of US data, and lastly, UNICEF Dataset. However, the dataset that I found on Kaggle suited the requirements for this project.

https://www.kaggle.com/vijaysimhan/student-admissions-data-for-a-university

3.2 Dataset Information

The dataset is around 857KB that includes 56 attributes. Some of the attributes include demographics, type of the student, courses, grades, parents' information, and the financial needs of students. The data includes many null values which will require using preprocessing techniques to be able to use the different machine learning and data mining algorithms.

3.2.1 Variable Dictionary

Variable Name	Description
STUDENT IDENTIFIER	Student Identifier
STDNT_AGE	Age of the Student Enrolled
STDNT_GENDER	Gender of the student
STDNT_BACKGROUND	Background of Student
IN_STATE_FLAG	Indicator of whether Student is in the same state as the university
INTERNATIONAL_STS	Indicator of whether Student is an International Student
STDNT_MAJOR	Student's Major course in university
STDNT_MINOR	Student's Minor course in university
STDNT_TEST_ENTRANCE1	Student's Entrance 1 score
STDNT_TEST_ENTRANCE2	Student's Entrance 2 score
STDNT_TEST_ENTRANCE_COMB	Student's score calculated both on Entrance1 & Entrance2 score
FIRST_TERM	First semester year
CORE_COURSE_NAME_1_F	Core course 1 opted in the First semester
CORE_COURSE_GRADE_1_F	Grade in Core course 1 opted in the First semester
CORE_COURSE_NAME_2_F	Core course 2 opted in the First semester
CORE_COURSE_GRADE_2_F	Grade in Core course 2 opted in the First semester
CORE_COURSE_NAME_3_F	Core course 3 opted in the First semester
CORE_COURSE_GRADE_3_F	Grade in Core course 3 opted in the First semester
CORE_COURSE_NAME_4_F	Core course 4 opted in the First semester
CORE_COURSE_GRADE_4_F	Grade in Core course 4 opted in the First semester
CORE_COURSE_NAME_5_F	Core course 5 opted in the First semester
CORE_COURSE_GRADE_5_F	Grade in Core course 5 opted in the First semester
CORE_COURSE_NAME_6_F	Core course 6 opted in the First semester
CORE_COURSE_GRADE_6_F	Grade in Core course 6 opted in the First semester
SECOND_TERM	Second semester year
CORE_COURSE_NAME_1_S	Core course 1 opted in the Second semester
CORE_COURSE_GRADE_1_S	Grade in Core course 1 opted in the Second semester
CORE_COURSE_NAME_2_S	Core course 2 opted in the Second semester
CORE_COURSE_GRADE_2_S	Grade in Core course 2 opted in the Second semester
CORE_COURSE_NAME_3_S	Core course 3 opted in the Second semester
CORE_COURSE_GRADE_3_S	Grade in Core course 3 opted in the Second semester
CORE_COURSE_NAME_4_S	Core course 4 opted in the Second semester
CORE_COURSE_GRADE_4_S	Grade in Core course 4 opted in the Second semester
CORE_COURSE_NAME_5_S	Core course 5 opted in the Second semester
CORE_COURSE_GRADE_5_S	Grade in Core course 5 opted in the Second semester
CORE_COURSE_NAME_6_S	Core course 6 opted in the Second semester
CORE_COURSE_GRADE_6_S	Grade in Core course 6 opted in the Second semester

HOUSING_STS	Indicator of whether the student is staying on campus or outside
RETURNED_2ND_YR	Indicates whether the student came back to the First semester in 2nd year
DISTANCE_FROM_HOME	Distance from the university to student's home
HIGH_SCHL_GPA	Student's High School GPA score
HIGH_SCHL_NAME	High School from where the student graduated
FATHER_HI_EDU_CD	Father's educational status code
FATHER_HI_EDU_DESC	Father's educational status
MOTHER_HI_EDU_CD	Mother's educational status code
MOTHER_HI_EDU_DESC	Mother's educational status
DEGREE_GROUP_CD	Degree code for which student has enrolled in university
DEGREE_GROUP_DESC	Degree for which student has enrolled in university
FIRST_TERM_ATTEMPT_HRS	# Hours attempted by a student (Or # Grade points attempted by Student in the First semester)
FIRST_TERM_EARNED_HRS	# Hours earned by a student (Or # Grade points earned by Student in the First semester)
SECOND_TERM_ATTEMPT_HRS	# Hours attempted by a student (Or # Grade points attempted by Student in the second semester)
SECOND_TERM_EARNED_HRS	# Hours earned by a student (Or # Grade points earned by Student in the second semester)
GROSS_FIN_NEED	Financial need of Student
COST_OF_ATTEND	Course Fees
EST_FAM_CONTRIBUTION	Estimated Family contribution towards course fees
UNMET_NEED	Unmet financial need of the student

Table 1 Variable Dictionary

Chapter 4- Project Analysis

4.1 Exploratory Data Analysis

Importing relevant libraries in Python to work with the dataset. This is the very first step in any data exploratory stage. This stage will cover the very basic information in the dataset such as how many attributes, how many null values, column names, as well as the summary of the dataset.

```
In [1]: import pandas as pd
        import numpy as np
        import matplotlib.pyplot as plt #plotting library in Python
        import seaborn as sns # used for data visualization and exploratory data analysis
```

Checking the features of the dataset, the total number of rows is 3400 and 56 attributes. The number of attributes is quite large compared to other large sets of data.

There are different types of attributes such as floats, integers, and objects within the whole dataset. It is also apparent that there are a lot of null values in every attribute, in this case, this needs to be preprocessed to have data with quality.

Some attributes such as core course names and grades have the largest percentage of null values, some reached 97% of null values. There are techniques to fill these null values that will be covered in the coming section.

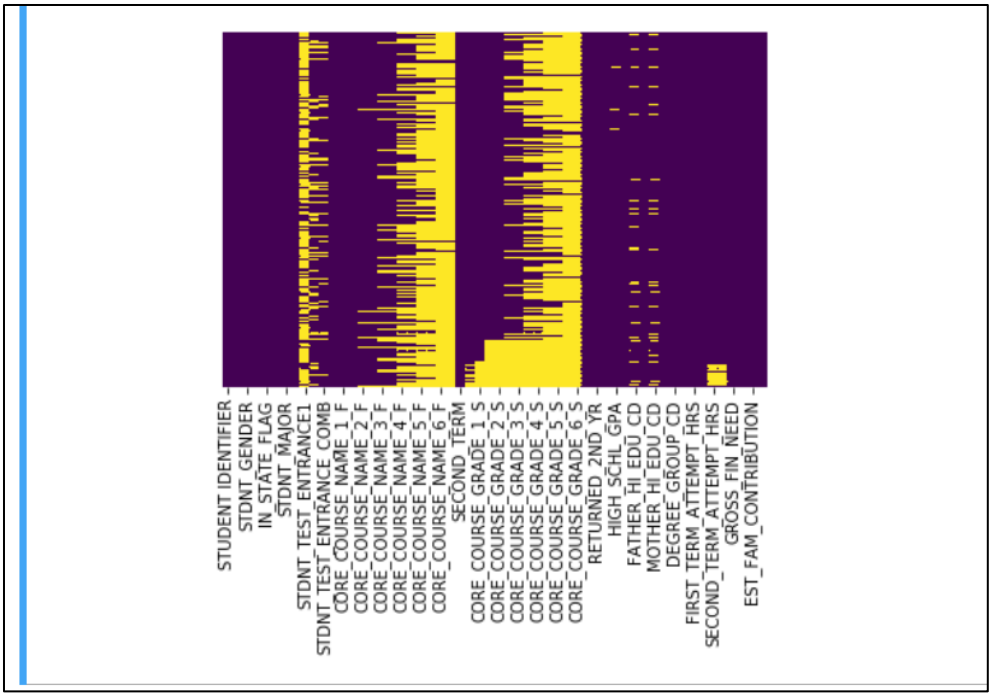


Figure 1 Heatmap Showing Null Values

The heat map above shows the highest attributes with null values visually. It shows that course names and grades have the highest null values fields than other attributes.

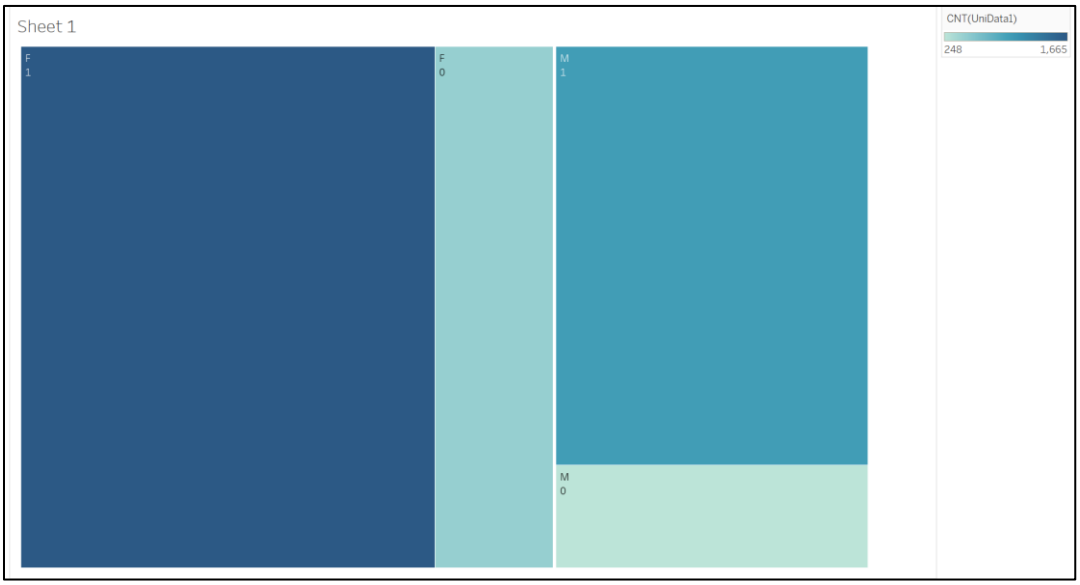


Figure 2 Returned Next Year Students

This plot from Tableau shows the number of students who returned to the university in the first semester in the 2nd year per gender. We can see that number of female students is more than the number of male students and, in the count, female student numbers who did not return is more than the number of male students.

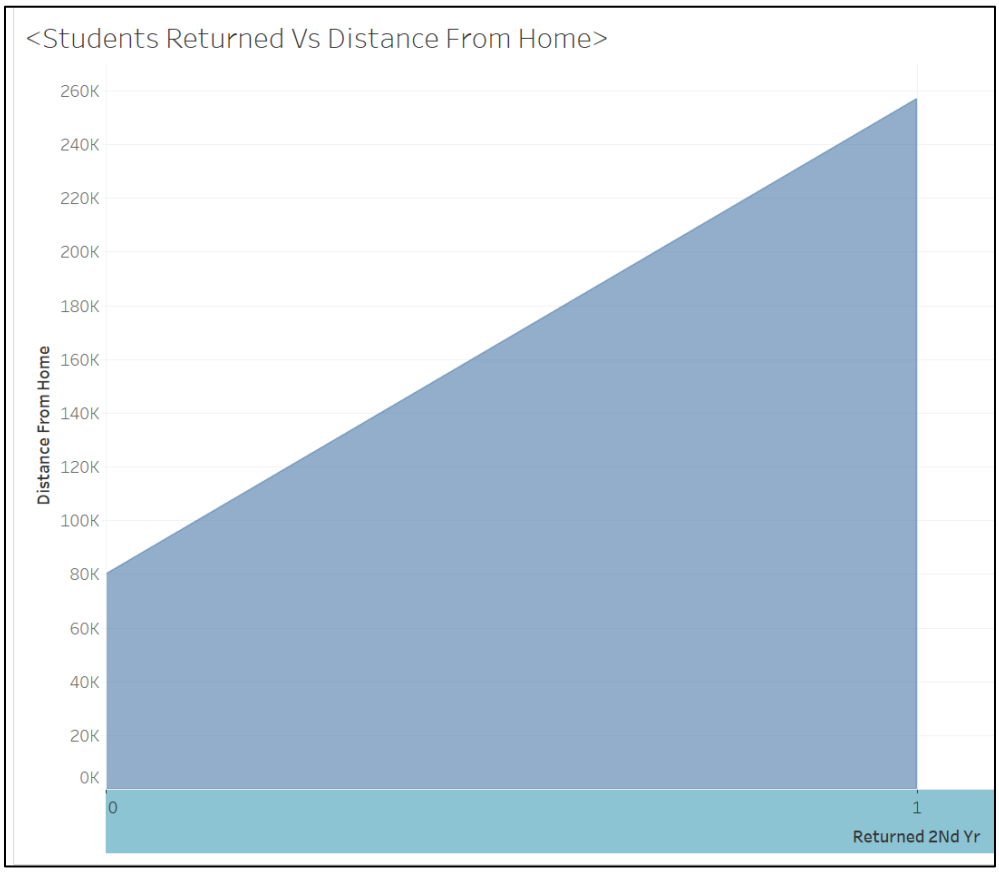


Figure 3 Students Returned Vs Distance from Home

One would think that the reason for students not returning in the first semester of the second year is due to transportation issues and distance from the university. However, this is not the case in this dataset. As it is shown, students who returned live far than those who did not return, but they could also be living in a hostel.

4.2 Data Quality Dimensions

To check data quality, we need to look at its six dimensions: Accuracy, Validity, Timelines, Completeness, Uniqueness, and Consistency. We need to take into consideration that not all dimensions might apply to one dataset.

- Accuracy, although the dataset is extracted through Kaggle, it falls under the Community Data License Agreement which makes it accurate as the data provider could be any entity or a person assigned from that entity.
- The validity, exploring the data showed that each type of variable is correct as the gender for example showed two groups, male and female.
- Completeness, there are a lot of missing values in the dataset which needs to be cleaned before we get insights
- Uniqueness, each row is represented with a unique identifier

4.3 Data Cleaning

As seen in the previous section of the exploratory analysis, there are a lot of null values found in the dataset. To be able to work with the data we have, it needs to be cleaned where the null values should be filled. In a categorical type of attribute, the null values will be filled with the mode of the attribute. However, for the continuous type of an attribute, the null values are filled with median if there was skewness. Whereas in an attribute where there is a normal distribution, the null values will be filled with a mean.

We saw in the heatmap, figure 1, that there was a huge distribution of null values in certain attributes than the others. Since those attributes will not give any additional information in our analysis, we'll drop them. The attributes dropped were: ['STDNT_TEST_ENTRANCE1',

'STDNT_TEST_ENTRANCE2', 'CORE_COURSE_NAME_4_F',

'CORE_COURSE_GRADE_4_F', 'CORE_COURSE_NAME_5_F',

'CORE_COURSE_GRADE_5_F', 'CORE_COURSE_NAME_6_F',

'CORE_COURSE_GRADE_6_F', 'CORE_COURSE_NAME_3_S',

'CORE_COURSE_GRADE_3_S', 'CORE_COURSE_NAME_4_S',

'CORE_COURSE_GRADE_4_S', 'CORE_COURSE_NAME_5_S',

'CORE_COURSE_GRADE_5_S', 'CORE_COURSE_NAME_6_S',

'CORE_COURSE_GRADE_6_S']

Now that we dropped some attributes, we have a total of 40 attributes that did not give us a good analysis. Attributes such as STUDENT IDENTIFIER, FATHER_HI_EDU_CD, MOTHER_HI_EDU_CD, and DEGREE_GROUP_CD can be dropped as well since these attributes are not adding any additional value to the dataset.

In the following, null values will be filled according to the attribute type and the skewness level.

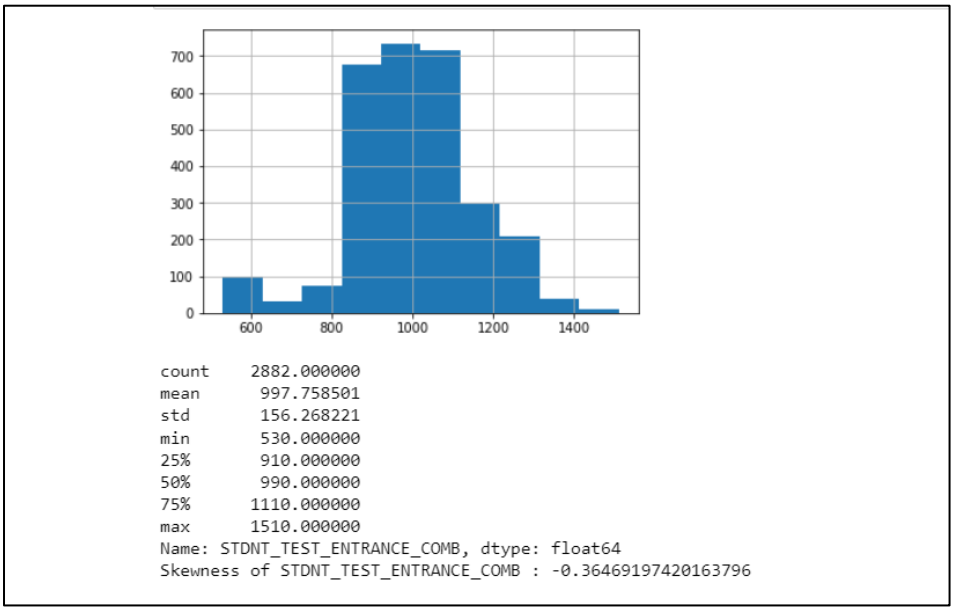


Figure 4 STDNT_TEST_ENTRANCE_COMB

The skewness of the above attribute is between -0.5 to 0.5, hence the data is normally distributed, and it can be filled with the mean.

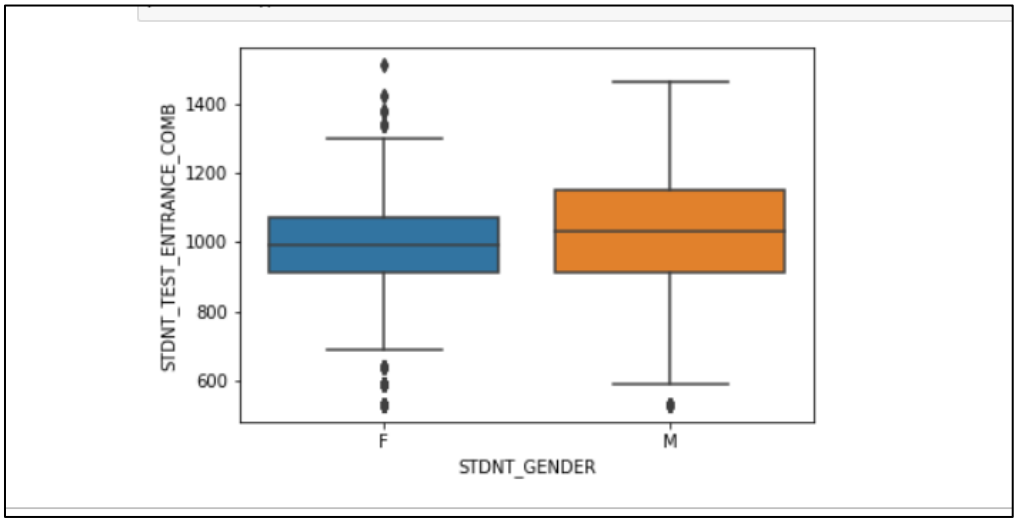
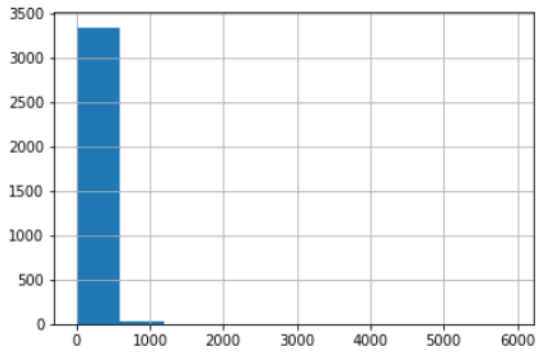


Figure 5 STDNT_TEST_ENTRANCE_COMB vs STDNT_GENDER

The above plot shows the distribution of the student test entrance scores in terms of gender, Male or Female. The average of STDNT_TEST_ENTRANCE_COMB varies significantly for males and females, hence, one of the practical ways to fill the null values is to do it separately for each gender.

The same shall apply to the other attributes with null values until we have a cleaned dataset, taking into consideration their type as highlighted at the beginning of this section.

```
18.439510065174744
STDNT_GENDER
F    101.694340
M     96.667729
Name: DISTANCE_FROM_HOME, dtype: float64
```




```
Out[25]: count    3375.000000
         mean       99.825185
         std       235.196984
         min         0.000000
         25%        69.000000
         50%        69.000000
         75%       138.000000
         max      5932.000000
         Name: DISTANCE_FROM_HOME, dtype: float64
```

```
In [26]: data_clean.DISTANCE_FROM_HOME.quantile(0.90)
Out[26]: 150.0
```

Figure 6 DISTANCE_FROM_HOME vs GENDER

One of the attributes still missing values is the DISTANCE_FROM_HOME which shows the distance between a student's home and the university. We need to cap the distance of students living away from the university to some decent value to obtain a normal distribution. Since the data type is float, then we'll check the skewness to fill the null values either by mean or median.

Through figure 7, it is apparent from the visualization that the attribute is highly skewed. Also, there is a huge gap between the 90th percentile which is 150 and the max value which is 5k. Hence, all points above the 90th percentile can be capped to the 90th percentile.

Out[27]:

STDNT_GENDER	count	mean	std	min	25%	50%	75%	max
F	2120.0	81.237264	52.071527	0.0	69.0	69.0	138.0	150.0
M	1255.0	76.185657	49.978015	0.0	69.0	69.0	136.0	150.0

Figure 7 DISTANCE_FROM_HOME vs GENDER

From the above table and skewness values, now mean can be used for missing value imputation as the data is normally distributed after treating the outliers. Also, there is a significant difference between the mean values of males and females, the best way to treat this is by using their mean values separately to fill the missing values.

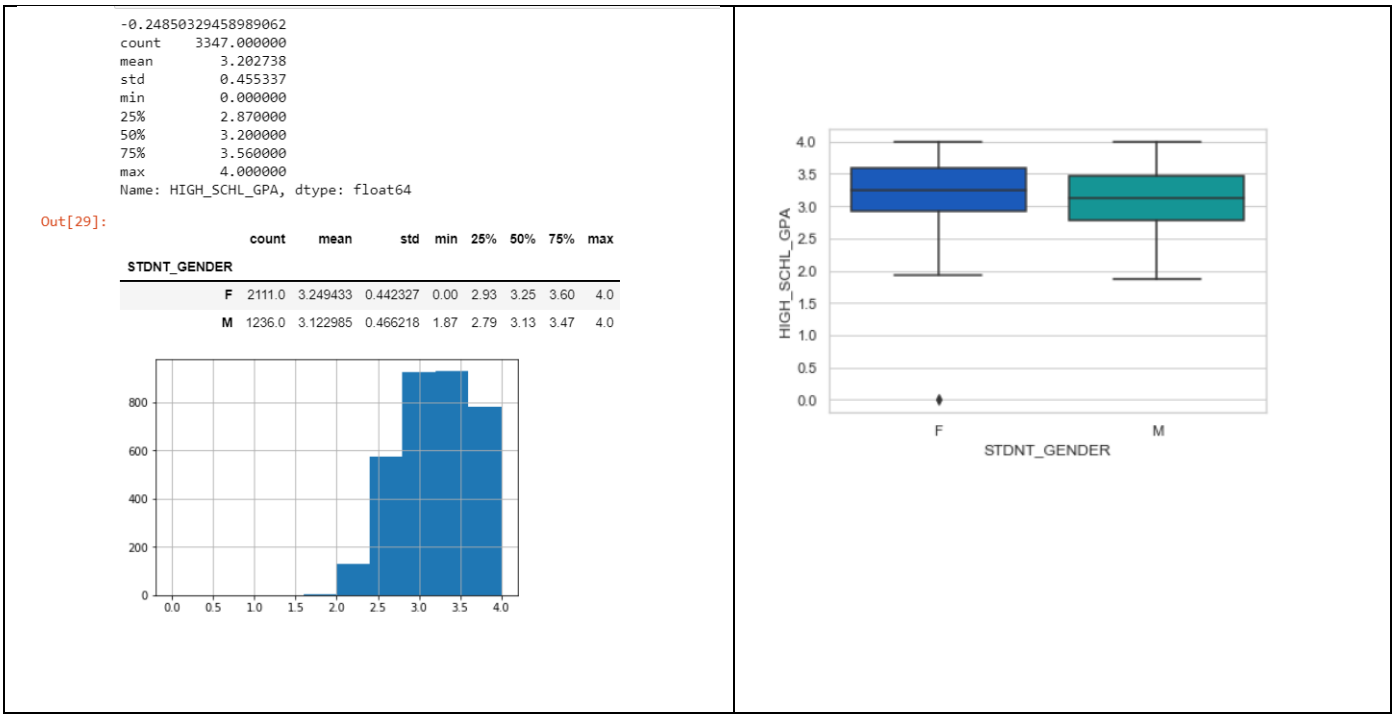


Figure 8 HIGH_SCHL_GPA

Looking at the HIGH_SCHL_GPA attribute, the variable is moderately right-skewed. Hence, using the median to fill the missing values. Again, we will follow the same strategy of using gender to fill respective missing values, fill the null values separately for each gender similar to what's been done above.

```
Out[33]: SCHOOL 11      249
         SCHOOL 130     245
         SCHOOL 10      238
         SCHOOL 1       209
         SCHOOL 2       129
                        ...
         SCHOOL 432       1
         SCHOOL 344       1
         SCHOOL 447       1
         SCHOOL 512       1
         SCHOOL 195       1
         Name: HIGH_SCHL_NAME, Length: 439, dtype: int64
```

Figure 9 HIGH_SCHL_NAME

Looking at the HIGH_SCHL_NAME attribute, we can see that it has a lot of categories i,e 439, however, there is no further information such as zone/state information available to group the schools and carry out further analysis of a particular school. Hence, this variable can be dropped as well.

Two new attributes can be derived from EARNED_HRS and ATTEMPT_HRS in respective terms for both terms, first and second which can be called FIRST_TERM_PERFORMANCE and SECOND_TERM_PERFORMANCE.

Out[37]:

	FIRST_TERM_PERFORMANCE	SECOND_TERM_PERFORMANCE
count	3400.000000	3191.000000
mean	0.872480	0.872315
std	0.189978	0.200743
min	0.000000	0.000000
25%	0.750000	0.769231
50%	1.000000	1.000000
75%	1.000000	1.000000
max	1.333333	1.055556

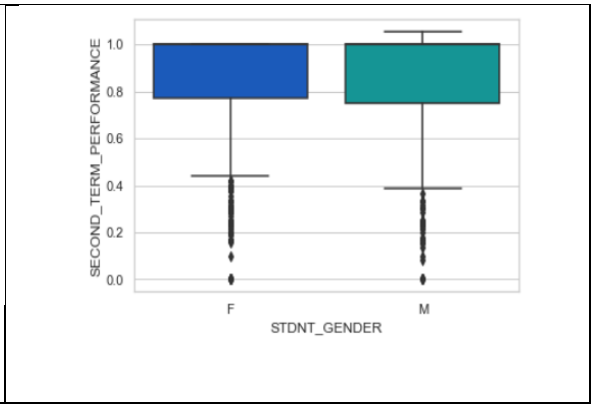


Figure 10 STUDENT Performance Details

By viewing the above figures, the mean can be used as the data is normally distributed. Also, since males and females have more or less the same mean value, we can fill the null values directly without doing it separately as done for the previous attributes.

Also, to avoid multicollinearity, the EARNED_HRS and ATTEMPT_HRS variables will be dropped since we derived two new features/attributes from them.

The target attribute for the project is RETURNED_2ND_YR where 1 represents returned, and 0 represents the contrary. It was best to swap between the values so that 1 represents the students who did not return since we aim to find out if students have dropped or not.

Compared to the first and second terms, the first term starts in August whereas the second starts in February. Hence, it is best to remove the month from our analysis.

It is best to remove the code number from each course's name as it is not adding any value in our analysis. When grouping student age with the target attribute, it was noticed that there's hardly any difference in student age between the two target classes (1 and 0).

Now that the dataset is completely cleaned, the next section will be used for visualizations to get insights.

4.4 Data Visualization

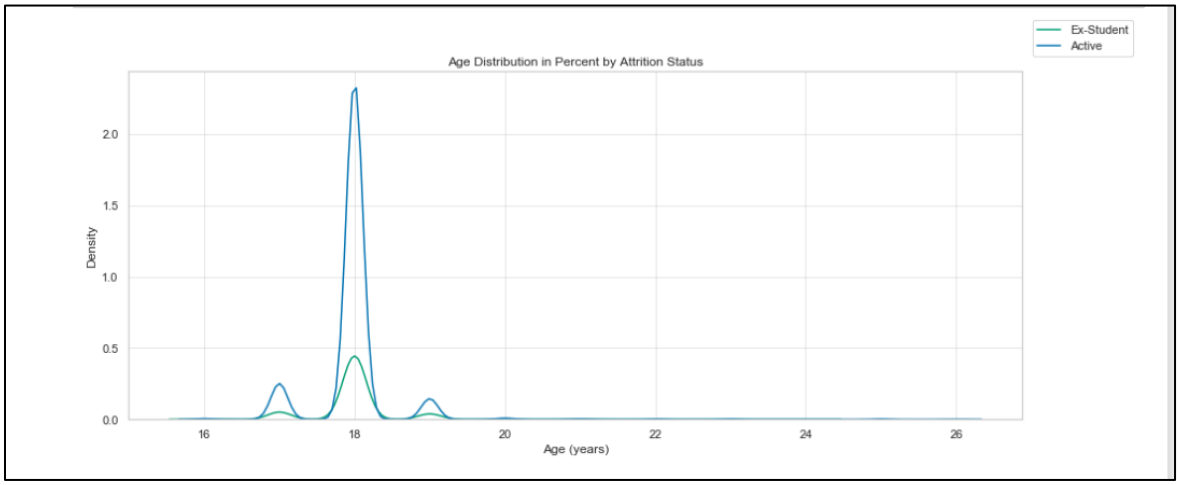


Figure 11 Attrition Status vs Student Age

We can see those students at 18 choose to drop the most where it is flat at ages 20 onwards.

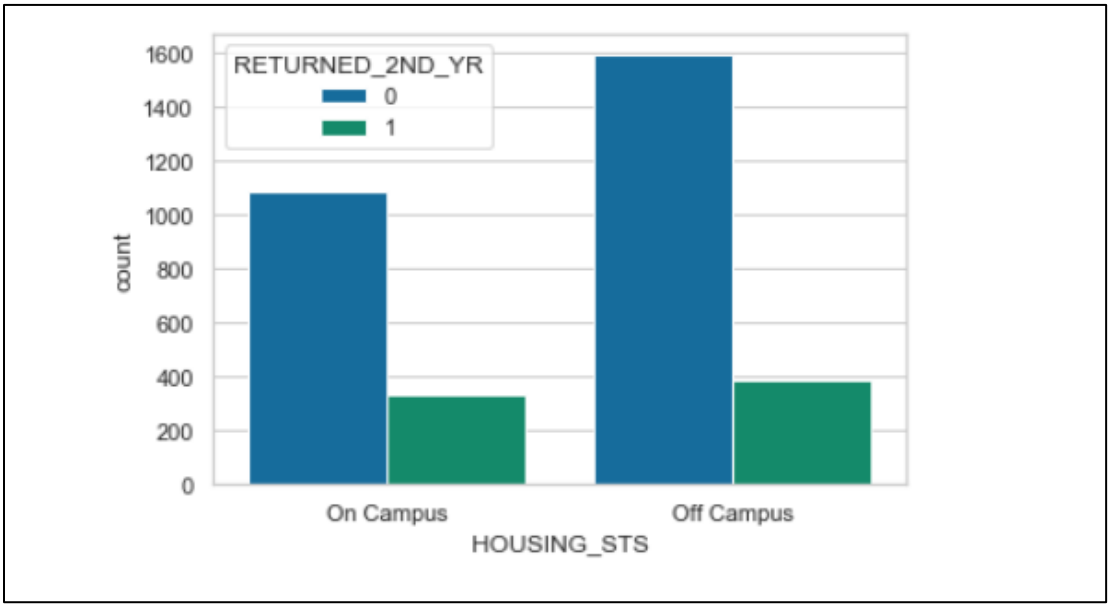


Figure 12 Housing Status compared to returning to campus or not

Students who dropped are represented in the green color bar. Those who are living off-campus have higher chances to drop than those who live on campus.

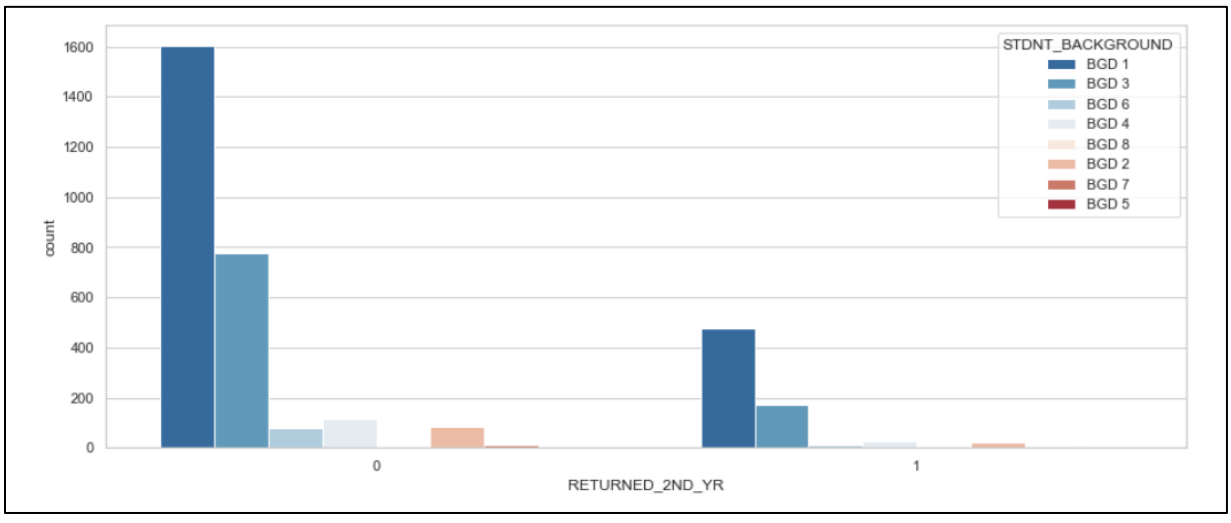


Figure 13 Students' Background

From the above plot, students from BGD1 have high attrition than other backgrounds where it reached more than 400+ students. Those are the same students who did not return the second year to the university.

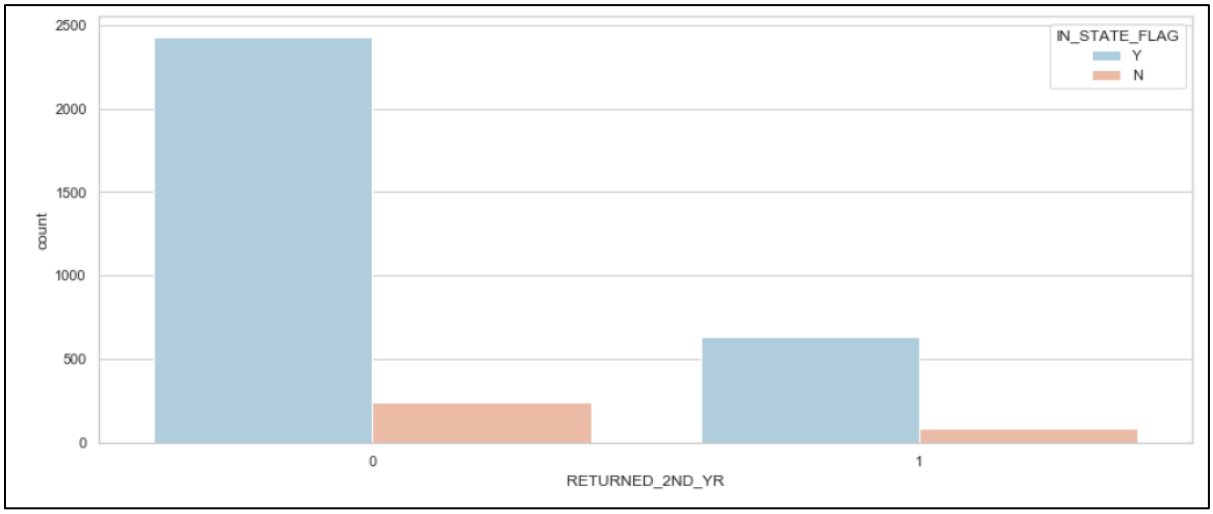


Figure 14 Same State Students

We can see that students who are in the same state of the university dropped more than those who are not in the same state. We can get from this insight that there could be other reasons that the distance from the university.

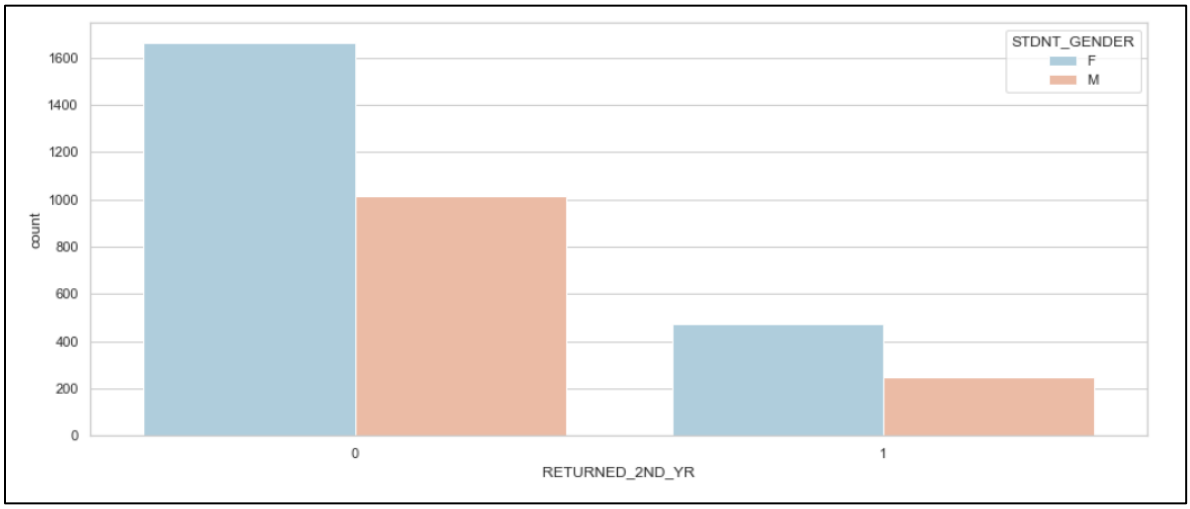


Figure 15 Student Gender vs Returned or not

Attrition among female students is higher than male counter partners. This means that the university needs to look at female students' needs and requirements to minimize the risk of their attrition.

4.5 Results – Exploratory Data Analysis

From the above exploratory plots, we can see that there are different insights we could get. Younger age plays a role in dropping from school, the exploratory analysis showed that those who aged 18 did not return to the university in the next semester whereas those who are 20 years old and onwards returned. Additionally, students who are living off-campus drop, however, the visualizations showed those who are living in the same state as the university drop more. Furthermore, students from BGD1 showed a higher drop amongst other backgrounds. It was not clear in the dataset which country or background this belongs to. Lastly, attrition among female students showed a higher percentage than the male.

4.6 Model Building

For the modeling, different algorithms were used. The main objective for using the models is to derive two things:

1) Optimum model with the highest sensitivity as it is the most that matter for this classification purpose

2) Feature importance i.e top features influencing student attrition

- **Scaling**: It is one of the most critical steps before creating any models in machine learning which can make a difference between a weak model and a better one. The most

common technique used in scaling is normalization which is used when we want to bound our values between two numbers either 0&1 or 1&-1. And standardization is used to have zero mean and a variance of 1. Used during the preprocessing of data.

- **Principal Component Analysis (PCA):** The main idea behind PCA is to reduce the dimensionality of the dataset that includes interrelated variables while keeping as much as possible the variations present in the data. Using the PCA will create new sets of uncorrelated variables that are called principal components which are ordered so that the first few ones retain the most variations in all the datasets. When using PCA, models become more efficient.
- **SMOTE**: Is identified as a powerful tool used for imbalanced data that represents an unequal number of classes in any classification problems where there is more representation of one class than the other. It stands for **Synthetic Minority Oversampling Technique.**
- **Train-test-split:** Used to split the dataset into training and evaluation separately. This technique is mostly used to evaluate the performance of a model in machine learning which can be used for classification and regression problems for any supervised learning algorithm.
- **Logistic regression**: This is a machine learning algorithm used in classification problems when the target variable is categorical and, in our case, to predict whether a student has returned or not next semester.
- **Random forest**: It is one of the successful machine learning algorithms due to its ability to provide good predictive performance, low overfitting, and interpretability.

- **Decision Trees**: Most widely used in supervised learning for classification (where target variable consists of discrete values) and regression problems (where the target variable consists of continuous values). They are constructed to split the dataset based on different conditions.
- **XGBoostClassifier**: e**X**treme **G**radient **B**oosting designed for tabular\structured datasets which are mainly used to enhance speed and model performance.
- **SVM**: Support Victor Machines is an algorithm that creates a line or a hyperplane that separates data into classes by using data as input and outputting a line that classes. Mainly used in data that has two classes in the target variable.

```
from sklearn.model_selection import train_test_split
from sklearn.pipeline import Pipeline,FeatureUnion
from sklearn.decomposition import PCA
from sklearn.preprocessing import StandardScaler

from sklearn.base import BaseEstimator, TransformerMixin
from sklearn.linear_model import LogisticRegression

# classification_report

from sklearn.metrics import classification_report, roc_auc_score,accuracy_score,precision_score,f1_score
from sklearn.metrics import confusion_matrix, precision_recall_curve, auc, roc_curve, recall_score

# sensitivity_specificity

from imblearn.metrics import sensitivity_specificity_support
from imblearn.over_sampling import SMOTE
from imblearn.combine import SMOTETomek
from imblearn.under_sampling import RandomUnderSampler

from sklearn.model_selection import GridSearchCV, StratifiedKFold, cross_val_score

from sklearn.metrics import confusion_matrix

from sklearn.ensemble import RandomForestClassifier, GradientBoostingClassifier

from sklearn import svm
```

Figure 16 Import the necessary libraries

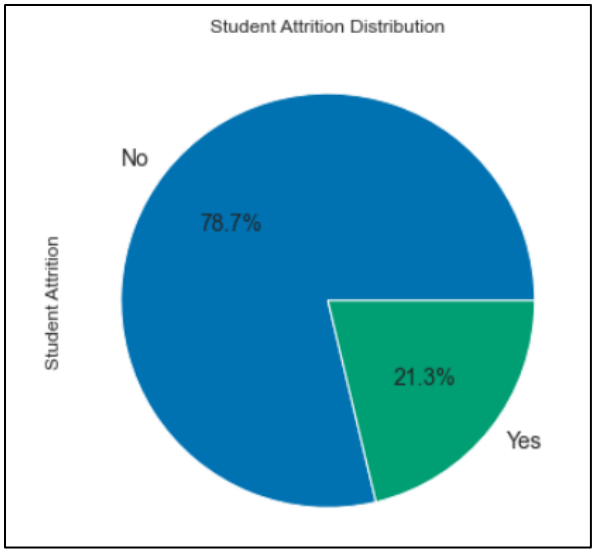


Figure 17 Imbalanced Data

In the above figure, it is obvious the class imbalance. To be able to work through class imbalance, we will use SMOTE as a class imbalance technique.

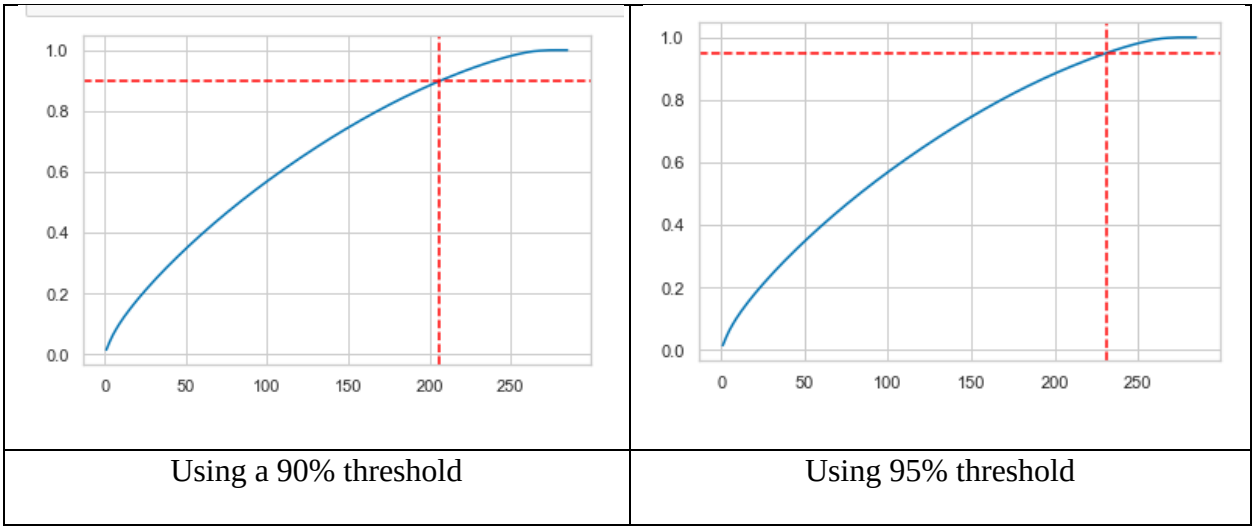
When using SMOTE alone for the imbalanced data, it might cause overfitting and bias during the training model where the class with the higher number will be chosen over the other class with a lower number of samples. Hence, it was best to use an undersampling technique such as SMOTETomek which is used as an undersampling method.

```
In [62]: # Create seperate train test dataset using SMOTE
         # let's divide data into train and test
         X = data_final.drop("STDNT_ATT", axis = 1)
         y = data_final.STDNT_ATT
         #oversample = SMOTE()
         #X_sm , y_sm = oversample.fit_resample(X, y)

         sm = SMOTETomek(random_state=42)
         X , y = sm.fit_resample(X, y)

         X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.3, random_state = 42)

In [63]: # let's check the shapes of train and test sets
         print(X_train.shape)
         print(y_train.shape)
         print(X_test.shape)
         print(y_test.shape)

         (3526, 285)
         (3526,)
         (1512, 285)
         (1512,)
```

Figure 18 Running the SMOTETomek for undersampling

We can see that the output is 3526 students records as a training set, whereas, 1512 records for the testing set. The 285 represents the number of features the data has.

After that, the PCA model is used to train the data with a threshold of 90% and 95%.



Using a 90% threshold	Using 95% threshold

Figure 19 Using the PCA model for training

When checking how many variables are required for the given threshold of variance, it was clear from the above figure that approx 206 components are required for a variance threshold of 90%, and approx. 231 components are required for the variance threshold of 95%. We shall consider the threshold as 90% and proceed in our study since neither 90% nor 95% has much of a difference between them.

When running the models, StandardScaler is used which can be defined as an operation that works in independent features which resize the distribution of the values in a variable.

Logistic Regression + PCA

```
Model parameters for Logistic Regression + PCA
Best AUC:  0.8176315622421126
Best hyperparameters:  {'logistic__C': 20, 'logistic__penalty': 'l2', 'pca__n_components': 231}
Confusion matrix - Test dataset :
[[685  79]
 [152 596]]
Train set results :
Sensitivity:    0.84
Specificity:    0.93
AUC of train data set:          0.88
Test set results :
Sensitivity:    0.8
Specificity:    0.9
AUC of test data set:           0.85
```

Figure 20 Logistic Regression + PCA output

The illustration above shows the scoring of the logistic regression with different model evaluation metrics. When running the code, the model created 440 fits of the total 44 candidates.

Random Forest + PCA

```
Fitting 5 folds for each of 720 candidates, totalling 3600 fits
Model parameters for Random Forest + PCA
Best AUC:  0.7673716877536405
Best hyperparameters:  {'model__criterion': 'entropy', 'model__max_depth': 2, 'model__max_features': 'auto', 'model__min_sample
s_leaf': 100, 'model__n_estimators': 100}
Confusion matrix - Test dataset :
[[715  49]
 [187 561]]
Train set results :
Sensitivity:    0.78
Specificity:    0.95
AUC of train data set:          0.86
Test set results :
Sensitivity:    0.75
Specificity:    0.94
AUC of test data set:           0.84
```

Figure 21 Random Forest + PCA output

With certain parameters, the random forest + PCA showed an output of 3600 of total fits.

XGBoostClassifier

```
[09:59:42] WARNING: C:/Users/Administrator/workspace/xgboost-win64_release_1.5.0/src/learner.cc:1115: Starting in XGBoost 1.3.
0, the default evaluation metric used with the objective 'binary:logistic' was changed from 'error' to 'logloss'. Explicitly se
t eval_metric if you'd like to restore the old behavior.
Model parameters for XGBoost with PCA
Best AUC:  0.8497923132012414
Best hyperparameters:  {'subsample': 0.8, 'n_estimators': 500, 'min_child_weight': 5, 'max_depth': 7, 'learning_rate': 0.25, 'g
amma': 1.5, 'colsample_bytree': 0.6}
Confusion matrix - Test dataset :
[[702  62]
 [138 610]]
Train set results :
Sensitivity:    0.97
Specificity:    1.0
AUC of train data set:          0.98
Test set results :
Sensitivity:    0.82
Specificity:    0.92
AUC of test data set:           0.87
```

Figure 22 XGBoostClassifier Model

For the above model, a total of 1500 fits have resulted, more than the logistic regression but less than the random forest.

Decision Tree classifier

```
Fitting 5 folds for each of 56 candidates, totalling 280 fits
Model parameters for Decision Tree with PCA
Best AUC:  0.8498050449590198
Best hyperparameters:  {'model__max_depth': 4, 'model__min_samples_leaf': 100}
Confusion matrix - Test dataset :
[[537 227]
 [130 618]]
Train set results :
Sensitivity:    0.85
Specificity:    0.69
AUC of train data set:          0.77
Test set results :
Sensitivity:    0.83
Specificity:    0.7
AUC of test data set:           0.76
```

Figure 23 Decision Tree Classifier

```
In [79]: # plotting tree with max_depth=3
         dot_data = StringIO()

         export_graphviz(dt_best['model'], out_file=dot_data, filled=True, rounded=True,
                         feature_names=X.columns,
                         class_names=['Not Attrited', "Attrited"])

         graph = pydotplus.graph_from_dot_data(dot_data.getvalue())
         Image(graph.create_png())
```

Out[79]:

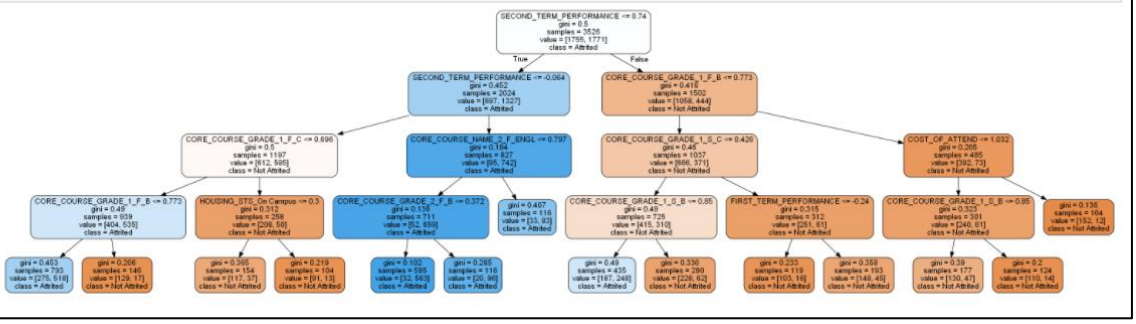


Figure 24 Decision Tree Plot

Random Forest without PCA

```
Fitting 5 folds for each of 720 candidates, totalling 3600 fits
Model parameters for Random Forest without PCA
Best AUC:  0.8170732871807115
Best hyperparameters:  {'model__criterion': 'entropy', 'model__max_depth': 10, 'model__max_features': 'auto', 'model__min_sampl
es_leaf': 5, 'model__n_estimators': 170}
Confusion matrix - Test dataset :
[[699  65]
 [145 603]]
Train set results :
Sensitivity:    0.85
Specificity:    0.94
AUC of train data set:          0.9
Test set results :
Sensitivity:    0.81
Specificity:    0.91
AUC of test data set:           0.86
```

Figure 25 Random Forest without PCA

In this model, PCA operation has been excluded. The final output of the model resulted in a total of 3600 fits.

4.7 Comparison of Different Models

Usually, for comparing the different classification models, data scientists most often use model accuracy. However, that might not be the case always.

There are other ways to compare the models which are used in this project. The different model parameter tools for comparison[1] used:

- **Sensitivity**: Another word is "Recall", is the ratio of true positives to total actual positives in the data
- **Specificity**: Ratio of true negatives to the total of negatives in the data
- **Accuracy**: Ration of correct predictions of total predictions
- **AUC**: The **A**rea **U**nder the **C**urve, measures the ability of a classifier to distinguish between classes, the higher, the better the performance
- **F1** Score: combines more than one comparison metric

Figure 26 shows the several models used and the output shows the important model evaluation parameters.

```
In [272]: consolidate_summary
```

Out[272]:

	Model	Accuracy - Train	Sensitivity - Train	Specificity - Train	AUC - Train	F1 - Train	Accuracy - Test	Sensitivity - Test	Specificity - Test	AUC - Test	F1 - Test
0	Logistic Regression + PCA	0.88	0.839074	0.928205	0.88	0.88	0.85	0.796791	0.896597	0.85	0.84
0	Random Forest + PCA	0.86	0.777527	0.949858	0.86	0.85	0.84	0.750000	0.935864	0.84	0.83
0	XGBoost with PCA	0.98	0.970638	0.997151	0.98	0.98	0.87	0.815508	0.918848	0.87	0.86
0	Decision Tree with PCA	0.77	0.851496	0.688319	0.77	0.79	0.76	0.826203	0.702880	0.76	0.78
0	Random Forest without PCA	0.90	0.852626	0.940171	0.90	0.89	0.86	0.806150	0.914921	0.86	0.85

Figure 26 A consolidation Summary of all the models and the evaluation metrics

[1] https://medium.com/analytics-vidhya/how-to-select-performance-metrics-for-classification-models-c847fe6b1ea3

The main focus would be to maximize sensitivity rather than accuracy as a university would like to correctly identify those students who are on the verge of dropping out and hence correct action can be applied. The higher the sensitivity, the better our case.

Considering only sensitivity, the below Models are best in order:

1. Decision Tree with PCA
2. XGBoost with PCA
3. Random Forest without PCA

However, if we also consider AUC and look for probable overfit, then 'Random Forest without PCA' outperforms the other two because of following reasons:

1. Decision Tree with PCA has the lowest AUC-Test.
2. XGBoost with PCA has 97% train sensitivity while only 81.5% test sensitivity which signifies probable overfit.
3. Random Forest without PCA has 86% AUC-test and all other parameters while compared with train values shows signs of the stable model.

Chapter 5 - Conclusion

5.1 Conclusion

Some universities are affected with high numbers of students' attrition due to the financial implications the attrition causes for these universities. Hence, analyzing students' data with machine learning algorithms help in reducing the attrition when getting the important features that affect students' attrition. Using a dataset from Kaggle, I have used different classifications models to predict students' attrition such as Logistic Regression, Random Forest, and Decision Tree Classifier. Using the model parameter comparisons criteria such as accuracy, sensitivity, and specificity, different percentages were derived from running the codes. As a result, Random Forest without PCA can be deployed into production to correctly identify the student's attrition. It will correctly identify student attrition with 80.6% confidence.

5.2 Recommendations

When getting the feature importance, there were 15 features identified using the random forest causing students' attrition.

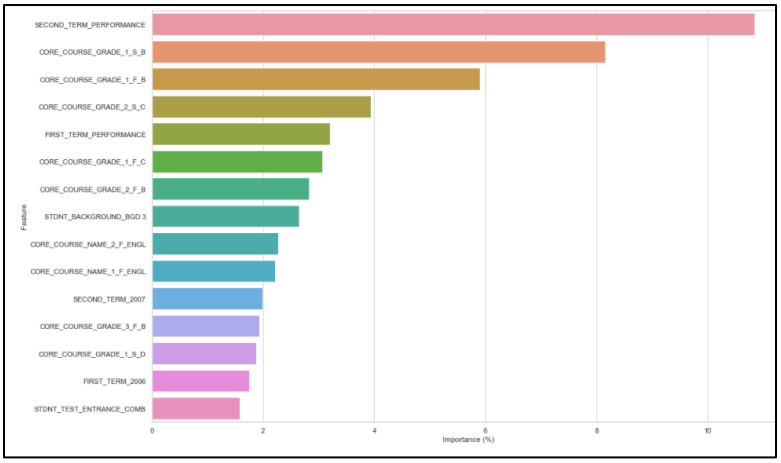


Figure 27 Feature importance that affect students' attrition

From using the random forest, we could see the most 15 variables that affect students' attrition.

```
SECOND_TERM_PERFORMANCE     -0.003942
CORE_COURSE_GRADE_1_S_B     -0.680222
CORE_COURSE_GRADE_1_F_B     -0.443693
CORE_COURSE_GRADE_2_S_C     -0.485831
FIRST_TERM_PERFORMANCE       0.062550
CORE_COURSE_GRADE_1_F_C     -0.275320
CORE_COURSE_GRADE_2_F_B     -0.151967
STDNT_BACKGROUND_BGD 3      -0.268182
CORE_COURSE_NAME_2_F_ENGL   -0.225702
CORE_COURSE_NAME_1_F_ENGL   -0.223245
SECOND_TERM_2007            -0.138133
CORE_COURSE_GRADE_3_F_B     -0.048540
CORE_COURSE_GRADE_1_S_D     -0.164264
FIRST_TERM_2006             -0.138133
STDNT_TEST_ENTRANCE_COMB    -0.046495
Name: 0, dtype: float64
```

Figure 28 Coefficient scores

From the above coefficients, the below suggestions can be given to universities or the Ministry of Education:

- Improve the teaching standards or grading for core course 1 opted in the first and second semester.
- Maybe more focus is given to core course 2 in the first and second semester because of which students can get good grades in these subjects and their grades are impacted in

other subjects. The possible idea is to investigate the teaching methodology of this subject so that it doesn't impact other subjects.

- Increase the students' enrollment in ENGL in the first and second semesters to reduce student attrition. Looks like a language barrier and with a common language, students from different language backgrounds can be made more interested in the teachings.
- Enroll more students with BGD 3 as those are less likely to drop and may positively influence others.
- Students enrolled in the first semester during 2006 and the second semester during 2007 shows fewer chances for attrition. Universities might have introduced some mechanism during these years that might have attracted students towards the studies. Further details can be discussed with relevant stakeholders to get more ideas on this.
- The performance indexes show that students with low performance in the first term are more prone to attrition.

5.3 Future Work

For future work, I would like if there were collaborations with the universities to share their real datasets about students instead of searching for an old dataset on the open sources data. In this way, universities with high attrition rates could benefit from different models to prevent student attritions.

Bibliography (APA format)

Abu-Oda, G. S., & El-Halees, A. M. (2015). Data mining in higher education: university student dropout case study. *International Journal of Data Mining & Knowledge Management Process*, *5*(1), 15.

Aljohani, O. (2016). Analyzing the Findings of the Saudi Research on Student Attrition in Higher Education. *International Education Studies*, *9*(8), 184-193. https://doi.org/10.5539/ies.v9n8p184

Aulck, L., Velagapudi, N., Blumenstock, J., & West, J. (2016). Predicting student dropout in higher education. *arXiv preprint arXiv:1606.06364*.

Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of educational data mining*, *1*(1), 3-17.

Berens, J., Schneider, K., Gortz, S., Oster, S., & Burghoff, J. (2019). Early Detection of Students at Risk - Predicting Student Dropouts Using Administrative Student Data from German Universities and Machine Learning Methods. *Journal of Educational Data Mining*, *11*(3), 1-41. https://doi.org/10.5281/zenodo.3594771

Dekker, G. W., Pechenizkiy, M., & Vleeshouwers, J. M. (2009). Predicting Students Drop Out: A Case Study. *International Working Group on Educational Data Mining*.

Grau-Valldosera, J., & Minguillón, J. (2014). Rethinking Dropout in Online Higher Education: The Case of the Universitat Oberta de Catalunya. *International Review of Research in Open and Distributed Learning*, *15*. https://doi.org/10.19173/irrodl.v15i1.1628

Johnson, N. (2012). The Institutional Costs of Student Attrition. Research Paper. *Delta Cost Project at American Institutes for Research*. Retrieved from: https://eric.ed.gov/?id=ED536126

Lee, S., & Chung, J. Y. (2019). The Machine Learning-Based Dropout Early Warning System for Improving the Performance of Dropout Prediction. *Applied Sciences*, *9*(15), 3093. https://doi:10.3390/app9153093

Maher, M., & Macallister, H. (2013). Retention and attrition of students in higher education: Challenges in modern times to what works. *Challenges in modern times to what works*, *3*(2), 62-73. http://doi.org/10.5539/hes.v3n2p62

Manhães, L. M. B., da Cruz, S. M. S., & Zimbrão, G. (2014). Evaluating performance and dropouts of undergraduates using educational data mining. In *Proceedings of the Twenty-Ninth Symposium on Applied Computing*.

Schneider, M. (2010). Finishing the first lap: The cost of first-year student attrition in America's four-year colleges and universities. Washington, DC: American Institutes for Research. Retrieved from http://www.air.org/files/AIR_Schneider_Finishing_the_First_Lap_Oct10.pdf

Shaw, M., Burrus, S. W.M., & Ferguson, K. (2016). *Factors That Influence Student Attrition in Online Courses*. Online Journal of Distance Learning Administration. https://www.researchgate.net/publication/308310140_Factors_that_Influence_Student_Attrition_in_Online_Courses

Yukselturk, E., Ozekes, S., & Türel, Y. K. (2014). Predicting Dropout Student: An Application of Data Mining Methods in an Online Education Program. *European Journal of Open, Distance and E-Learning*, *17*(1), 118–133. https://doi.org/10.2478/eurodl-2014-0008